IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 6, PP 14-21

Mapping Hindi Text Document Into Graph Structure

Yogita Shelar[1], Pragyamani Sharma[2], Vaishali Salvi[3], Jyothi Arun[4]

[1]*(Information Technology, University Of Mumbai, India)*
[2]*(Information Technology, University Of Mumbai, India)*
[3]*(Information Technology, University Of Mumbai, India)*
[4]*(Information Technology, University Of Mumbai, India)*

Abstract: *In this paper, we present a Mapping Hindi text document into graph structure and external program for pre- processed frequently, proper or further. It is derived using natural language processing as follows. Firstly subject – verb – object triplets are automatically extracted from the Hindi dictionary obtained for each sentence in the document. Secondly, extracting the triplets and enhanced by linking them to their corresponding co-referenced named entity, by resolving pronominal anaphors as well as attaching the associated Word Net synset.*

Keywords: *Pre-processed, frequently, atomically, pronominal, anaphors.*

I. Introduction

This project describes the Document summarization refers to the task of creating document surrogates that are smaller in size but retain various characteristics of the original document. To automate the process of abstracting, researchers generally rely on a two phase process. First, key textual elements, e.g., keywords, clauses, sentences, or paragraphs are extracted from text using linguistic and statistical analyses. In the second step, the extracted text may be used as a summary. Such summaries are referred to as 'extracts'. Alternatively, textual elements can be used to generate new text, similar to the human authored abstract. Summarization of Hindi documents contains historical information is also plays as important role for students and teachers who want to read a large number of documents related to history. Summarization system helps them to read and learn the shorter version of overall complete document Summarization system helps them to read and learn the shorter version of overall complete document.

Automatic Text Summarization is an important and challenging area of Natural Language Processing. The task of a text summarizer is to produce a synopsis of any document or a set of documents submitted to it. Analysis of Text-Documents has been an active area of research for the past few years. It involves extensive use of Natural Language Processing techniques for analysing semantic structures of the text. Semantic analysis of a document means to analyse the meaning or transitions in meaning of the sentences or of different clauses and the relation among them. There are a number of approaches to semantic analysis. Semantic analysis can be done at the sentence level, the paragraph level, or even at the text level on different languages.

Hindi is the official and the most widely spoken language in India. As of pronoun resolution, for the pronouns having more than one possible antecedent, the pronoun resolution mechanism of this system captures the ambiguity. The approach discussed here is to perform semantic analysis at the sentence level where the Hindi text is scanned for pronouns and the corresponding referents resolved.

Summaries differ in several ways. A summary can be an extract i.e. certain portions (sentences or phrases) of the text is lifted and reproduced verb, where as producing an abstract involves breaking down of the text into a number of different key ideas, fusion of specific ideas to get more general ones, and then generation of new sentences dealing with these new general ideas.

We propose interactive system with advance features programming approach to analyse the Text Document based on semantic graphs derived from subject-object-verb triplets. This technique applies for providing documents and produces their graph description.

These properties make natural language processing a challenging task but also an interesting research topic especially with the increasing use of information technology. Many computer applications that involve natural

language such as machine translation, question answering and information extraction are dependent on modeling natural language in some way.

II. Problem Definition

The Semantic graphs are intractable for large domains, and they do not represent performance or meta-knowledge very well. Some properties are not easily expressed using a semantic graph, e.g., negation, disjunction, and general non-taxonomic knowledge.

I optimized the best solution that analysing the Hindi text document using Processing Parser for triplet extraction, creating graph, normalization for large document.

III. Scope

The There are a number of component where construction for Graph description with features are very useful. For example, an information retrieval system could present an automatically built summary in its list of retrieval results, for the user to quickly decide which documents are interesting and worth opening for a closer look this is what Google models to some degree with the snippets shown in its search results. Other example is that these information for government officials, businessmen, researches, Neural Networks, Fuzzy Logic, etc. and Hence these documents contain a huge amount of information which needs to be summarized so that reader can read and learn the important information from these documents with ease.

IV. Review Of Literature

Various methods have been proposed to achieve extractive Analysing. Most of them are based on scoring of the sentences.

Dr.Latesh Malik, et. al.[1], Discussed single document summarization using extraction method for Hindi text, which uses statistical and linguistic features. It uses Hindi Wordnet to tag appropriate POS of word for checking SOV of the sentences which uses sixstatistical and two linguistic features. It also uses genetic algorithm to optimize the summary generated based on the text feature terms with less redundancy.

Ibrahim F. Moawad, et. al.[2], Described a novel approach is presented to create an abstractive summary for a single document using a rich semantic graph reducing technique. The approach summaries the input document by creating a semantic graph called Rich Semantic Graph for the original document, reducing the generated semantic graph to more abstracted graph, and generating the abstractive summary from the reduced graph but in English.

Agarwal, et. al.[3], Proposed the algorithm for anaphora resolution has been tested extensively. The accuracy of anaphora resolution is 96% for simple sentence not for original document and complex sentences; the accuracy is of the order of 80%. This method works by searching sentences in the text that are semantically related thorough anaphors, analyzing their semantic s and exploiting the latter for s resolving respective anaphors.

Ng Choon-Ching, et. al.[4], Proposed an existing need for text summarizers that small devices like PDA has emerged the development of text summarization of web pages. Authors have identified problems for text summarization in several areas such as dynamic content of web pages, diverse summary definition of text, and different benchmark of evaluation measurements. Besides, authors also found advantages of certain methods that increased the accuracy of web page classification. In the future work, author plan to investigate machine learning techniques to incorporate additional features for the improvement of text summarization quality. The additional features authors are currently considering include linguistic features such as discourse structure, lexical chains, semantic features such as name entities, time, location information etc

Visual Gupta, et. al.[5], Describe the Punjabi text extractive system which consist of two phases 1) Pre Processing 2) Processing. In this paper term pre processing is defined as the phase which identify the word boundary, sentence boundary, Punjabi stop words elimination etc. and the processing phase sentence features are calculated and a weight is assigned to each sentence on the reference of which unwanted sentences are eliminated from the input text. It is described that the author tested the proposed system over fifty Punjabi news documents (with 6185 sentences and 72689 words) from punjabi ajit news paper and fifty punjabi stories (with 17538 sentences and178400 words). Accuracy of the system is varies from 81% to 92 %.

Niladri Chatterjee, et. al.[6], Described summarization technique for text document exploiting the semantic similarity between sentences to remove the redundancy from the text. It uses Random Indexing for compute the semantic similarity scores of sentences and graph-based ranking algorithms have been employed to produce an extract of the given text. The important is that the problem of high dimensionality of the semantic space corresponding text should be tackled by random indexing which is less expensive in computations and memory consumption and it included a training algorithm using Random Indexing which has to construct the Word space on complied text document so that resolve the ambiguities such as more efficiency.

M. C. Padma, et. al.[7], In a multi-script multi-lingual environment, a document may contain text lines in more than one script/language forms. It is necessary to identify different script regions of the document in order to feed the document to the OCRs of individual language. With this context, this paper proposes to develop a homothetic algorithmic model to identify and separate text lines Telugu, Hindi and English scripts from a printed multilingual document. The proposed method uses the distinct features of the target script and searches Preparation of Papers for the International Journal of Engineering and Science www.theijes.com The IJES Page 4 for the text lines that possess the anticipated features. Experimentation conducted involved 1500 text lines for learning and 900 text lines for testing. The performance has turned out to be 98.5%.

Erika Velazquez-Garcia, et. al.[8], Proposed A novel method to organize, search and display groups of document according to topics they contain based on the collection of synonyms, and hyponyms, hyponyms of each terms Thus, each user would have a personalized and dynamic organized of documents thereby it takes more time for text processing. Sunil Kumar, et. al.[9], Suggested a novel scheme for the extraction of textual areas of an image using globally matched wavelet filters. A clustering-based technique has been devised for estimating globally matched wavelet filters using a collection of ground truth images. We have extended our text extraction scheme for the segmentation of document images into text, background, and picture components (which include graphics and continuous tone images). Multiple, two-class Fisher classifiers have been used for this purpose. We also exploit contextual information by using a Markov random field formulation-based pixel labeling scheme for refinement of the segmentation results. Experimental results have established effectiveness of our approach..

M. Swamy Das, et. al.[10], Described document should be composed of text contents in different languages in multilingual country. It is necessary to identify the language region of the document before feeding the document to the related OCR system. Advantage of this paper is that a model to identify script type of different text portions using visual clues. Here seven features are covered, such as, bottom max row, top horizontal lines, vertical lines, bottom component, tick component and top holes, and bottom holes have been used to identify the script document. Identification accuracy of above 93%is achieved.

Patricia J. Crossno, et. al.[11], Proposed Topic-View, that is LSA concepts provide good summarizations overbroad groups of documents, while LDA topics are focused on smaller groups. LDA's limited document groups and its probabilistic mechanism for determining a topic's top terms support better labeling for document clusters than LSA concepts, but the document relationships defined by the LSA model do not include extraneous connections between disparate topics identified by LDA.

Futoshi Iwama, et. al.[12], Described a novel framework for creating a parser to process and analyze texts written in a "partially structured" natural language. They designed a framework for combinatorial text-parsers and implemented the text-parser combination system. Those system can combine existing natural language parsers with formal language parsers by using novel combinatory and then generate text- parsers applying specific natural language processing to specified parts of entire text-documents.

Timo Honkela, et. al.[13], Described a novel method, Grounded Inter-subjective Concept Analysis (GICA), for the analysis and visualization of individual differences in language use and conceptualization. Joakim Nivre, et. al.[14], Described the application of Malt-Parser, a transition-based dependency parser, to three Indian languages–Bangla, Hindi and Telugu – in the context of the NLP Tools Contest at ICON 2009. In the final evaluation, Malt-Parser was ranked second among the participating systems and achieved an unlabeled attachment score close to 90% for Bangla and Hindi, and over 85% for Telugu, while the labeled attachment score was 15–25 percentage points lower. It is likely that the high unlabeled accuracy is achieved thanks to a relatively low syntactic complexity in the data sets, while the low labeled accuracy is due to the limited amounts of training data.

We concluded from all the papers, Hindi is the official and the most widely spoken language in India. We discussed various methods for summarization. But many of techniques are found on English and other languages but very few methods on Hindi text document. Summarization of Hindi documents contains historical information is also plays as important role for students and teachers who want to read a large number of documents related to history. Summarization can be two types: 1. Extractive Summarization 2. Abstractive Summarization. In both extractive and abstractive summarization technique rule based approach can be used in which various handcrafted rules are to be created on the basis of which summary of the text document can be generated.

V. Theory of the idea

The semantic graph is utilized in order to represent the document's semantic structure. Our approach is based on the research presented in and further developed in. While in semantic graph generation was relying on the proprietary NLP Win linguistic tool for deep syntactic analysis and pronominal reference resolution, we take advantage of the co-referenced named entities as well as the triplets extracted from Hindi Word Net and derive rules for pronominal anaphora resolution and graph generation. For generating the graph, triplets are first linked to their associated named entity (if appropriate). Furthermore, they are assigned their corresponding Word Net synset. We obtain a directed semantic graph, the direction being from the subject node to the object node, and the connecting link (or relation) is represented by the predicate.

The semantic graph is obtained after processing the input document and passing it through a series of sequential operations composing a pipeline (see Figure 1):

There are six Modules:

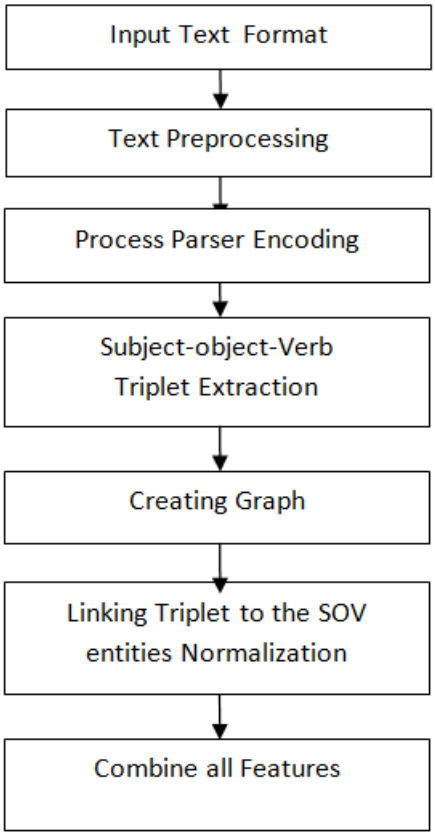

Fig.1

Basic details about these modules as below:

1. Document Visualization Process is start with the original document, a program that processing its input data to produce output that is used as input to another program. The output is said to be a **preprocessed** form of the input data, which is often used by some subsequent programs like compilers, which is using

Malt Parser and found the names entities from the document and split between the three different entity typed: people, location and organization. Hence gives the output in this format.

Input – विपुल ने रमेश को खत लिखा।

Output -

1.	1	विपुल	विपुल	NN	6	k1
2.	2	ने	ने	PSP:ने	1	lwg__psp
3.	3	रमेश	रमेश	NN	6	k4
4.	4	को	को	PSP:को	3	lwg__psp
5.	5	खत	खत	NN	6	pof
6.	6	लिखा	लिखा	VM	0	main
7.	7	।	।	.	6	rsym

Fig.2

2. Process Parser Encoding - In this step processing is same as above with extensible. Here Triplets are enhanced by resolving anaphors for subsets of pronouns and their objective, reflexive and possessive forms as well as the relative pronoun. It included extra accuracy.

1.	1	विपुल	विपुल	NN	6	k1	subject	1	7	5	0.4	0.2
2.	3	रमेश	रमेश	NN	6	k4	object	1	7	5	0.4	0.2
3.	5	खत	खत	NN	6	pof	object	1	7	2	0.4	0.4
4.	6	लिखा	लिखा	VM	0	main	verb	1	7	1	0.4	0.8
5.	।											

Fig.3

3. Subject-Object-Verb Triplet Extraction - Here Triplets are extracted from each sentence independently, without text outside of the sentence. Triplets are linked to their corresponding co-referenced named entity. We apply the algorithm for obtaining triplets from a Process-Parser output described in the output. They are highlighted differently as shown below output.

1. 1 विपुल विपुल NN 6 k1 subject 1 7 5 0.4 0.2
2. 6 लिखा लिखा VM 0 main verb 1 7 1 0.4 0.8

3. 3 रमेश रमेश NN 6 k4 object 1 7 5 0.4 0.2
4. 1 विपुल विपुल NN 6 k1 subject 1 7 5 0.4 0.2
5. 6 लिखा लिखा VM 0 main verb 1 7 1 0.4 0.8
6. 5 खत खत NN 6 pof object 1 7 2 0.4 0.4

Fig.4

4. Crating Graph, In this step generation system components were evaluated by comparing their output. Here I included features word-position, position-tag, word-depth, dependency-tag, subject-object, line-no, sentence-len, word-freq, sentence-similarity , term-frequency-inverse-sentence-frequency.

1. 1 विपुल विपुल NN 6 k1 subject 1 7 5 0.4 0.2 0.1 0.3 0.5 1 5 4
2. 6 लिखा लिखा VM 0 main verb 1 7 1 0.4 0.8
3. 3 रमेश रमेश NN 6 k4 object 1 7 5 0.4 0.2 0.0 0.6 0.5 2 2
4. 1 विपुल विपुल NN 6 k1 subject 1 7 5 0.4 0.2 0.1 0.3 0.5 1 5 4
5. 6 लिखा लिखा VM 0 main verb 1 7 1 0.4 0.8
6. 5 खत खत NN 6 pof object 1 7 2 0.4 0.4 0.0 0.3 0.3 1 0 0

Fig.5

5 Normalization - In this step the process of organizing the attributes and tables of a relational database to minimize data redundancy. Here we included extra features : word-position, position-tag, word-depth, dependency-tag, subject-object-verb Tag, line-no, sentence-len, word-freq, sentence-similarity , term-frequency-inverse-sentence-frequency, Page Rank, Authority, Hub, In-Links, Out-Links, Next-Neigh; output is in given format as below.

1. 1 विपुल 0.0 1 0 1 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0.0 1.0 1.0 0.0 1.0 0.0 1.0 0.0 1.0 0.8 0.0

2. 1 लिखा 1.0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 0 0 1 0.0 0.0 1.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

3. 1 रमेश 0.4 1 0 0 0 0 0 0 0 0 0 1 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 0 0.0 1.0 1.0 0.0 0.0 1.0 1.0 1.0 0.4 1.0 0.0

4. 1 विपुल 0.0 1 0 1 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0.0 1.0 1.0 0.0 1.0 0.0 1.0 0.0 1.0 0.8 0.0

5. 1 लिखा 1.0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 0 0 1 0.0 0.0 1.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

6. 1 खत 0.8 1 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

1 0 0 1 0 0.0 0.3 1.0 0.3 0.0 0.0 0.0 0.0 0.0 0.0 0.0

Fig.6

6. Combine Features - In this step combines multiple features within a layer into one based on a common attribute value which is used for SVM, Nuclear power center.

I. 0 विपुल लिखा रमेश 0.0 1 0 1 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 1 0 0 0.0 1.0 1.0 0.0 1.0 0.0 1.0 0.0 1.0 0.8

0.0 1.0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

1 0 0 1 0.0 0.0 1.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.4 1

0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1

0 0.0 1.0 1.0 0.0 0.0 1.0 1.0 1.0 0.4 1.0 0.0

II. 1 विपुल लिखा खत 0.0 1 0 1 0 0 0 0 0 0 0 0

0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0

0 0 0 0 1 0 0 0.0 1.0 1.0 0.0 1.0 0.0 1.0 0.0 1.0

0.8

0.0 1.0 0 1 0 0 0 0 0 0 0 0 0 0 0 0

0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0

1 0 0 1 0.0 0.0 1.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.8 1

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1

0 0.0 0.3 1.0 0.3 0.0 0.0 0.0 0.0 0.0 0.0 0.0

Fig.7

VI. Conclusions

Hindi is the official and the most widely spoken language in India. In this project, I proposed Hindi text documents visualization technique based on semantic graphs. This techniques can be applied only to the original document, each of the system components were detailed, processing with own parser generation for triplet extraction, created graph and normalization, then merge all the features, they were used for giving the answer in detail form.

I can concluded that the presented output helpful for the user, However, more experiments evaluating, researches, Neural Networks, etc

VII. Future Work

As future work, we intend to conduct additional research in order to extending the system adding several components , such as another triplet extraction and enhance the semantic representation as well as document summary. There has been a huge effort in developing effective and efficient methods for finding named entities and extracting triplet from text document and linking for creating graph and then merge all the features. Because of that, any gain (even very small ones) may represent a relevant saving in time and computational resources.

References

Journal Papers:

[1]. Dr.Latesh Malik,"Test Model for Summarizing Hindi Text using Extraction Method",(*Proceedings of 2013 IEEE Conference on Information and Communication Technologies) (ICT 2013).*

[2]. Ibrahim F. Moawad, Information Systems Dept.Faculty of Computer and Information Sciences "Semantic Graph Reduction b Approach for Abstractive Text Summarization",(*Ain shams University Cairo, Egypt ibrahim_moawad@cis.asu.edu.eg 2012 IEEE).*

[3]. Sachin AGARWAL Manaj SRIVASTAVA, "Anaphora Resolution in Hindi Documents",(*Indian Institute of Information Technology – Allahabad Allahabad, UP, India2007IEEE)*

[4]. Do Phuc, University of Information Technology,"Using SOM based Graph Clustering for Extracting *Main Ideas from Documents",(VNU-HCM HoChiMinh City,VietNam phucdo@uit.edu.vn 2008 IEEE)*.

[5]. Vishal Gupta and Gurpreet Singh Lehal*," Automatic Punjabi Text Extractive Summarization system." Proceedings of COLING 2012: Demonstration Papers, pages 199–206, COLING 2012, Mumbai, December 2012.*

[6]. Niladri Chatterjee,"Extraction-Based Single-Document Summarization Using Random Indexing",(*19th IEEE International Conference on Tools with Artificial Intelligence IEEE2007).*

[7]. M. C. Padma, P. A. Vijaya, "Monothetic Separation of Telugu, Hindi and English Text Lines from a Multi Script Document", (*Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA - October 2009)*

[8]. Erika Velazquez-Garcia, Ivan Lopez-Arevalo, Victor Jesus Sosa-Sosa Information Technology, Laboratory CINVESTAV – Tamaulipa,"Representing Document Semantics by Means of Graphs", (*http://www.google.com visited in September 2011)*.

[9]. Sunil Kumar, Rajat Gupta, Nitin Khanna, Student Member, IEEE, Santanu Chaudhury, and Shiv Dutt Joshi, "Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model",(*IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 16, NO. 8, AUGUST 2007).*

[10]. M. Swamy Das, D. Sandhya Rani, C R K Reddy, "Heuristic based Script Identification from Multilingual Text Documents", International Conf. On Recect Advances in Information Technology (*RAIT-2012).*

[11]. Patricia J. Crossno, Andrew T. Wilson and Timothy M. Shead, "TopicView:VisuallyComparingTopicModelsofTextCollections", Scalable Analysis and Visualization Sandia National Laboratories Albuquerque, NM 87185 USA {pjcross, atwilso, tshead}@sandia.gov

[12]. Futoshi Iwama, Taiga Nakamura, Hironori Takeuchi, "Constructing Parser for Industrial Software Specifications Containing Formal and Natural Language Description",IBM Research - Tokyo IBM Japan Yamato, Kanagawa, Japan gamma@jp.ibm.com taiga@jp.ibm.com hironori@jp.ibm.com

[13]. Timo Honkela, Juha Raitio, Krista Lagus and Ilari T. Nieminen, "Subjects on Objects in Contexts: Using GICA Method to Quantify Epistemological Subjectivity",Aalto University School of Science Dep't of Information and Computer Science P.O.Box 15400, FI-00076 AALTO, Finland Email: first.last@aalto.fi

[14]. JoakimNivre, "Parsing Indian Languages with MaltParser", Uppsala University Department of Linguistics and Philology E-mail: joakim.nivre@lingfil.uu.se.